

# X-Linked Genes and Risk of Orofacial Clefts: Evidence from Two Population-Based Studies in Scandinavia

Astanand Jugessur<sup>1,2,\*</sup>, Øivind Skare<sup>1,3,9</sup>, Rolv T. Lie<sup>3,4</sup>, Allen J. Wilcox<sup>5</sup>, Kaare Christensen<sup>6,7,8</sup>, Lene Christiansen<sup>6</sup>, Truc Trung Nguyen<sup>4</sup>, Jeffrey C. Murray<sup>9</sup>, Håkon K. Gjessing<sup>1,3</sup>

**1** Division of Epidemiology, Norwegian Institute of Public Health, Oslo, Norway, **2** Craniofacial Research, Murdoch Childrens Research Institute, Royal Children's Hospital, Parkville, Australia, **3** Department of Public Health and Primary Health Care, University of Bergen, Bergen, Norway, **4** Medical Birth Registry of Norway, Norwegian Institute of Public Health, Bergen, Norway, **5** Epidemiology Branch, National Institute of Environmental Health Sciences (National Institute of Health, Durham, North Carolina, United States of America), **6** Department of Epidemiology, University of Southern Denmark, Odense, Denmark, **7** Department of Clinical Biochemistry and Pharmacology, Odense University Hospital, Odense, Denmark, **8** Department of Clinical Genetics, Odense University Hospital, Odense, Denmark, **9** Departments of Pediatrics, Epidemiology and Biological Sciences, University of Iowa, Iowa City, Iowa, United States of America

## Abstract

**Background:** Orofacial clefts are common birth defects of complex etiology, with an excess of males among babies with cleft lip and palate, and an excess of females among those with cleft palate only. Although genes on the X chromosome have been implicated in clefting, there has been no association analysis of X-linked markers.

**Methodology/Principal Findings:** We added new functionalities in the HAPLIN statistical software to enable association analysis of X-linked markers and an exploration of various causal scenarios relevant to orofacial clefts. Genotypes for 48 SNPs in 18 candidate genes on the X chromosome were analyzed in two population-based samples from Scandinavia (562 Norwegian and 235 Danish case-parent triads). For haplotype analysis, we used a sliding-window approach and assessed isolated cleft lip with or without cleft palate (iCL/P) separately from isolated cleft palate only (iCPO). We tested three statistical models in HAPLIN, allowing for: i) the same relative risk in males and females, ii) sex-specific relative risks, and iii) X-inactivation in females. We found weak but consistent associations with the oral-facial-digital syndrome 1 (*OFD1*) gene (formerly known as *CXORF5*) in the Danish iCL/P samples across all models, but not in the Norwegian iCL/P samples. In sex-specific analyses, the association with *OFD1* was in male cases only. No analyses showed associations with iCPO in either the Norwegian or the Danish sample.

**Conclusions:** The association of *OFD1* with iCL/P is plausible given the biological relevance of this gene. However, the lack of replication in the Norwegian samples highlights the need to verify these preliminary findings in other large datasets. More generally, the novel analytic methods presented here are widely applicable to investigations of the role of X-linked genes in complex traits.

**Citation:** Jugessur A, Skare Ø, Lie RT, Wilcox AJ, Christensen K, et al. (2012) X-Linked Genes and Risk of Orofacial Clefts: Evidence from Two Population-Based Studies in Scandinavia. PLoS ONE 7(6): e39240. doi:10.1371/journal.pone.0039240

**Editor:** Devendra Amre, University of Montreal, Canada

**Received:** February 27, 2012; **Accepted:** May 17, 2012; **Published:** June 19, 2012

**Copyright:** © 2012 Jugessur et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** Center for Inherited Disease Research (CIDR) is fully funded through a federal contract from the National Institutes of Health (NIH) to The Johns Hopkins University, Contract Number N01-HG-65403. This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences; by NIH grants DE08559, P60 DE13076, NIH P30 ES05605, and RO1 DE-11948-04; and by the Norwegian Research Council (NFR 177522). The authors also thank the United States National Institute of Dental and Craniofacial Research for underwriting a significant proportion of the genotyping costs by CIDR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: astanand.jugessur@fhi.no

These authors contributed equally to this work.

## Introduction

Orofacial clefts are relatively common craniofacial birth defects, with a birth prevalence of about 1–2/1000. They require extensive surgical, nutritional, dental, speech, behavioral and medical interventions, and thus impose a substantial economic and personal health burden [1,2]. As with other complex traits, multiple genetic and environmental risk factors are thought to underlie these birth defects [3].

Clefts are characterized by a particularly strong genetic component, as evidenced by studies of familial recurrence risk and heritability [4]. First-degree relatives of an affected individual

have a 30–40 fold higher recurrence risk compared with the background population [5,6], and heritability estimates were recently reported to be as high as 91% for CL/P and 90% for CP in a large Danish twin study [7].

The strong genetic component to clefting has spurred decade-long efforts to identify the genes underpinning these complex birth defects. Increased collaborative efforts coupled with major advances in high-density SNP genotyping arrays have heralded a new era of gene discovery for complex traits. For clefts, the first genome-wide association study (GWAS) identified a strong signal on chromosome 8q24 in individuals of central European ancestry. This association was subsequently replicated in three

**Table 1.** Review of family-based methods for association analysis of X-chromosome markers.

Reference	Method	Extended name	Attributes
Ho and Bailey-Wilson [23]	X-TDT	X-linkage transmission/disequilibrium test (TDT)	This is a TDT for linkage on the X chromosome in the presence of linkage disequilibrium (LD). Under $H_0$ of no linkage between disease and marker, the number of transmissions of the variant allele in $n$ pairs of heterozygous mothers and their affected children has a binomial distribution with mean $n/2$ and variance $n/4$ . The test statistic is a Z-score with a continuity correction, and $H_0$ is rejected if Z departs significantly from 0. As with TDT, X-TDT is readily extended to allow the analysis of phenotypically discordant sib pairs if parental genotypes are unavailable (suitable for late-onset diseases). It can also combine sib-pair and case-parent triad analysis to enhance statistical power.
Horvath et al. [21]; Knapp [22]	XS-TDT; XRC-TDT	X-linked sibling TDT; Reconstruction-combined TDT for X-chromosome markers	As X-TDT above, these are tests for linkage between an X-chromosomal marker and a disease in the presence of LD. XS-TDT uses the genotypes of discordant sibships if genotypes are not available from the parents. It divides the siblings into same-sex groups to account for a possible male/female difference in disease prevalence. XRC-TDT reconstructs parental genotypes from the genotypes of their offspring and corrects for bias that arise from the reconstruction. Data from families in which parental genotypes are available are combined with families in which genotypes of unaffected sib pairs are available.
Ding et al. [24]	XPDT; XMCPDT	X-chromosomal pedigree disequilibrium test; Monte Carlo pedigree disequilibrium test for X-linked markers	XPDT tests for LD in the presence of linkage. It can be applied to any pedigree structure. XMCPDT is an extension of XPDT and infers missing parental genotypes using a Monte Carlo sampling approach. XPDT is limited to same-sex discordant sib pairs when parental data are missing, resulting in lower statistical power. XMCPDT on the other hand requires allele frequency estimates to compensate for missing parental genotypes. XMCPDT appears to have superior power than XSTDT, XRCTDT or XPDT when there are missing data, but Type 1 errors can be inflated when a large proportion of parental genotypes are missing.
Chung et al. [25]	X-APL	A modification of the "association in the presence of linkage test (APL)" that accommodates X-chromosome markers	Like XPDT, X-APL can use singleton or multiplex families. The APL statistic is based on difference between the observed versus the expected number of a specific allele in affected siblings conditional on the parents' genotypes. X-APL infers missing parental genotypes in linkage regions by using identity-by-descent (IBD) parameters for affected siblings. X-APL can test individual markers or haplotypes. For haplotype tests, X-APL assumes no recombination between the markers within the families in the sample, and the EM algorithm is used for haplotype phase estimation. X-APL can also perform sex-stratified analyses to account for different penetrance of disease in males versus females.
Zhang et al. [27]	X-LRT	A likelihood ratio test of association for X-linked markers.	X-LRT is a likelihood-based method and enables estimation of genetic risks. The method is designed for singleton families but can also allow additional siblings. Missing parental genotypes can be accounted for using the EM algorithm, and even more efficiently using sibling genotype information when available. For haplotype relative risk estimation, X-LRT assumes no recombination between markers, parental mating to be random, and haplotype penetrance to be multiplicative for females. For sex-specific analysis, separate risk parameters are introduced for males and females in single-marker analyses, but in haplotype analyses the data are divided into two sets, one containing only male cases and the other only female cases.
This paper	HAPLIN	A full likelihood model for haplotype associations at autosomal and X-linked markers.	HAPLIN is a likelihood-based method and enables estimation of genetic risk associated with marker haplotypes both for autosomal and X-linked markers. It applies to case-parent triad data, possibly combined with independent controls and/or complete control-parent triads. Missing data are imputed using the EM algorithm. On the X chromosome, HAPLIN provides a range of model options depending on haplotype effects in females versus males. A complete sex stratification implies different haplotype frequencies, different baseline risks and different relative risks between males and females. Alternatively, haplotype frequencies can be assumed equal, as can haplotype relative risks. The risk response pattern may depend on the number of risk haplotypes, and X-inactivation in females can be incorporated.

doi:10.1371/journal.pone.0039240.t001

independent GWAS [8–10]. In addition to the 8q24 locus, these studies also identified significant associations with several genes, including v-maf musculoaponeurotic fibrosarcoma oncogene homolog B (*MAFB*), ATP-binding cassette sub-family A member 4 (*ABCA4*), ventral anterior homeobox 1 (*VAX1*), paired box 7

(*PAX7*) and interferon regulatory factor 6 (*IRF6*) [4]. *IRF6* is particularly noteworthy, being the only gene on this list to be confirmed as a major player for clefts through approaches other than GWAS [11–13].

The above studies and virtually all association studies of clefts (as well as other complex traits) have targeted primarily autosomal markers, without attention to potential contributions of X-linked common gene variants. This is partly because most of the statistical methods originally designed for association analysis were only targeted towards the analysis of autosomal markers. The finding that X-linked gene variants may be implicated in a number of complex traits [14–19] has encouraged the development of statistical methods for analyzing X-linked markers. The majority of these methods are extensions of the transmission/disequilibrium test (TDT) [20], for example, the X-linked sibling TDT (XS-TDT) [21], the reconstruction-combined TDT for X-chromosome markers (XRC-TDT) [22],

**Table 2.** Sample distribution according to cleft type, sex, and population.

Cleft type	Norway		Denmark	
	Males	Females	Males	Females
iCL/P	202	109	114	52
iCPO	54	60	33	36

doi:10.1371/journal.pone.0039240.t002

the X-linkage TDT (X-TDT) [23], and the X-chromosome pedigree disequilibrium test (XPDT) [24]. Two additional tests compare observed versus expected distributions of a specific allele or haplotype in affected siblings, conditional on the parental genotypes. These are the 1) association in the presence of linkage (APL) test that accommodates X-chromosome markers (X-APL) [25], and 2) X-linked quantitative trait loci linkage mapping (X-QTL) [26]. Despite several attractive attributes of these methods (summarized in **Table 1**), an important limitation is that they can only provide a p-value for association, but not estimates of genetic risk. The exception is the likelihood ratio test (LRT) developed by Zhang and co-workers (X-LRT) [27].

An exploration of X-linked variants is particularly relevant when a complex trait is more common in one sex – as is seen for the two main types of orofacial clefts. For this study, we implemented new functionalities in the HAPLIN software [28] to enable X-chromosome marker analysis and an estimation of relative risks associated with either a single or double dose of an allele or haplotype. We considered various model parameterizations that address a range of causal scenarios relevant to an X-linked disease locus. This included allowing for different baseline risks for males and females (to reflect the higher prevalence of CL/P in males), and accounting for X-inactivation in females (where one of the two copies of the X chromosome is inactivated in each cell to ensure similar gene dosage across the two sexes [29]).

We applied this new method to a collection of 48 SNPs in 18 cleft candidate genes on the X chromosome and used data from two national cleft studies in Scandinavia – one of the largest collections of orofacial cleft triads available. To our knowledge, no previous study has explored the role of X-linked genes in the etiology of orofacial clefts using association.

## Materials and Methods

### Study Populations, Candidate Genes and SNPs

From a population-based case-control study of orofacial clefts in Norway (1996–2001), 311 iCL/P and 114 iCPO case-parent triads were available for the current analysis. As a replication sample, we had a further 166 iCL/P and 69 iCPO case-parent

triads from a population-based study of orofacial clefts in Denmark (1991–2001). The sample distribution according to cleft type, sex, and population is provided in **Table 2**. Details regarding study design and participants have been provided elsewhere [30,31]. The 18 X-linked genes and 48 SNPs for the current analysis derive from a larger candidate-gene based study in which we examined 357 candidate genes in the same study populations [32]. A detailed description of these 18 X-linked genes and 48 SNPs is provided in the online **Table S1**.

### Statistical Analysis

**The HAPLIN software.** The statistical software HAPLIN [28] was specifically designed to analyze genetic and environmental risk factors in offspring-parent triads and case-control collections. It is based on log-linear modeling as originally described [33–41] and implements a full maximum-likelihood model for estimation. HAPLIN computes explicit estimates of relative risks with asymptotic standard errors and confidence intervals. It uses the expectation-maximization (EM) algorithm to impute genotypes that are missing at “random” (e.g. due to failed genotyping) and those missing by “design” (e.g. if DNA from a family member was not available for genotyping). The EM algorithm can also reconstruct unknown haplotype phase for haplotype analysis, but on the X chromosome this is not needed since phase can be deduced directly when data are non-missing.

Central to HAPLIN is a generalized linear model (glm) being estimated from the observed genotype frequencies—the M-step of the EM algorithm. The E-step consists of all three imputations described above, performed in a single step. The algorithm then alternates between the M-step and E-step until convergence is achieved. The results obtained from the EM algorithm correspond to the maximum-likelihood estimates of the model, which include gene frequencies and all relative risk parameters. However, to obtain the correct standard errors, confidence intervals and likelihood ratio test (LRT) for the models, HAPLIN corrects for the fact that imputation has taken place. If the imputed data were used uncorrected, they would seem to contain more information than what is actually available in the raw data.

**X-linked haplotype analysis using HAPLIN.** HAPLIN allows a range of X-chromosome models to be estimated, depending on assumptions made about allele effects in males versus females. The following models, summarized in **Table 3**, are examples of risk parameterizations provided by HAPLIN:

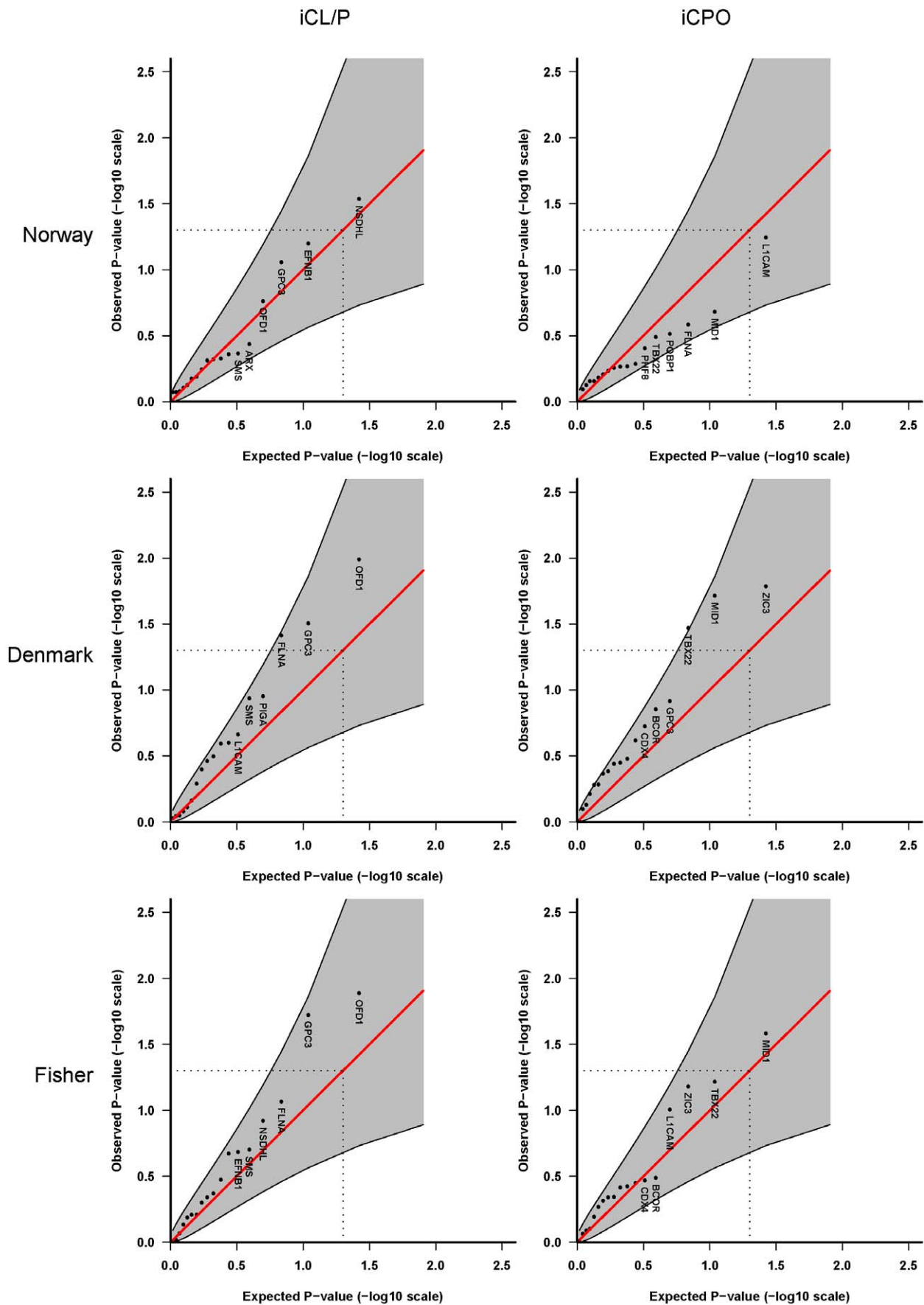
- **Model 1:** A shared baseline risk for males and females; a shared relative risk for males and females; no X-inactivation; a multiplicative dose-response relationship in females (1 risk parameter to be estimated).
- **Model 2:** Different baseline risks for males and females; a shared relative risk for males and females; no X-inactivation; a multiplicative dose-response relationship in females (2 risk parameters to be estimated)
- **Model 3:** Different baseline risks for males and females; different relative risks for males and females; no X-inactivation; a multiplicative dose-response relationship in females (3 risk parameters to be estimated)
- **Model 4:** Different baseline risks for males and females; a shared relative risk for males and females; X-inactivation; a multiplicative dose-response relationship in females (2 risk parameters to be estimated)
- **Model 5:** Different baseline risks for males and females; different relative risks for males and females; no assumption of multiplicative risks. We refer to this as the “free model”, with

**Table 3.** Assorted parameterization models for analysis of X-linked gene variants using the HAPLIN software.

Model	Male case		Female case		
	X <sub>1</sub>	X <sub>2</sub>	X <sub>1</sub> X <sub>1</sub>	X <sub>1</sub> X <sub>2</sub>	X <sub>2</sub> X <sub>2</sub>
Model 1	B	B*RR	B	B*RR	B*RR <sup>2</sup>
Model 2	B <sub>M</sub>	B <sub>M</sub> *RR	B <sub>F</sub>	B <sub>F</sub> *RR	B <sub>F</sub> *RR <sup>2</sup>
Model 3	B <sub>M</sub>	B <sub>M</sub> *RR <sub>M</sub>	B <sub>F</sub>	B <sub>F</sub> *RR <sub>F</sub>	B <sub>F</sub> *RR <sub>F</sub> <sup>2</sup>
Model 4	B <sub>M</sub>	B <sub>M</sub> *RR	B <sub>F</sub>	1/2*B <sub>F</sub> *(1+RR)	B <sub>F</sub> *RR
Model 5	B <sub>M</sub>	B <sub>M</sub> *RR <sub>M</sub>	B <sub>F</sub>	B <sub>F</sub> *RR <sub>F1</sub>	B <sub>F</sub> *RR <sub>F2</sub>

X<sub>1</sub> denotes the common allele and X<sub>2</sub> the variant or target allele for a given SNP; ‘\*’ denotes the product term; B represents the shared baseline risk for males and females; B<sub>M</sub> is the baseline risk for males only; B<sub>F</sub> is the baseline risk for females only; RR is the shared relative risk for males and females; RR<sub>M</sub> is the relative risk for males only; and RR<sub>F</sub> is the relative risk for females only. In Model 4, the risk for an X<sub>1</sub>X<sub>2</sub> female is the average of the two homozygotes; i.e. (B<sub>F</sub>+B<sub>F</sub>\*RR)/2 = B<sub>F</sub>(1+RR)/2. As this is not a log-linear model, HAPLIN replaces the heterozygous risk with B<sub>F</sub>√RR, i.e. the geometric mean of the two homozygous risks. Models 3 and 5 can be estimated assuming equal or unequal haplotype frequencies between males and females.

doi:10.1371/journal.pone.0039240.t003





**Figure 1. Single-marker analyses of 48 SNPs in 18 X-linked cleft candidate genes.** These analyses are based on **Model 2** in which we assume different baseline risks for males and females, a shared relative risk for males and females, and no X-inactivation. Quantile-quantile (QQ) plots of p-values for iCL/P (left-hand side) and iCPO (right-hand side). Top panels: Norwegian and Danish samples, respectively. Bottommost panels: Fisher combined p-values. Shaded areas represent 95% confidence interval bands and dotted lines indicate the expected ranked p-value of 0.05. Note that the oral-facial-digital syndrome 1 gene (*OFD1*) was formerly known as *CXORF5*.  
doi:10.1371/journal.pone.0039240.g001

the highest number of parameters to be estimated (4 risk parameters).

With fewer parameters, more assumptions are needed but power is improved (provided the model is correct). The log-linear model implemented in HAPLIN extends the X-LRT approach described by Zhang and colleagues [27], which essentially estimates **Model 5** for single SNP markers, although with a different parameterization. Even though stratifying the case-parent triads by sex may reduce the statistical power to detect an association, we performed sex-specific analyses to verify whether there is a stronger gene effect in one sex versus the other. **Model 1**, with the assumption of equal baselines, is less relevant to our data. **Models 3** and **5** can be estimated assuming equal allele frequencies between males and females, or by completely stratifying on sex. All models have natural extensions from single SNP markers to testing multiple haplotypes.

**Sliding windows and multiple testing.** Our study comprised one of the largest collections of case-parent triads for clefts from two population-based sources. Even so, it is unlikely that gene-effects are large enough that a single gene would remain significant after a full Bonferroni correction, even when the two sexes are analyzed together. The stringent requirement of ensuring an overall Type 1 error rate of  $\leq 5\%$  will be overly conservative, especially in a study such as this one where the candidate genes had been selected *a priori* for their potential roles in clefting.

To adequately deal with the multiple-testing issue, we followed a two-part approach. First, all p-values computed from single SNPs or haplotypes within a gene were summarized into a single p-value for that gene, corrected for within-gene multiple testing. Second, the single gene p-values were plotted together in a quantile-quantile (QQ) plot, which would reveal p-values more significant than what would be expected by chance given the number of genes being tested.

For the first part, HAPLIN includes the function `haplinSlide` (for details, see <http://www.uib.no/smis/gjessing/genetics/software/haplin/> or the HAPLIN help pages in **R**), which automates the analysis of a long sequence of single SNPs, or alternatively a sequence of overlapping sliding-windows with haplotypes. Overlapping sliding-windows will in principle increase the chance of “bracketing” a causal variant by having a haplotype with SNPs on each side of the variant. However, estimating haplotypes entails a certain loss of power due to the higher number of alleles taken into account and the unknown phase of the haplotypes. It is *a priori* not obvious whether a single-SNP approach or a sliding-window haplotype approach would have the best chance of detecting an association; therefore, we performed both single-marker and haplotype analyses on the current data. We restricted HAPLIN to use up to 4 SNPs in a sliding-window haplotype analysis, which typically ensures bracketing causal variants with two SNPs on each side. Using more than 4 SNPs in a haplotype would most likely entail an unnecessary loss of power. With longer haplotypes, the number of possible haplotypes increases exponentially, and any given haplotype will be found in very few individuals (if any), making effect estimation difficult. After running a sliding-window analysis, the results were summarized in the form of a single, overall p-value associated with each gene. This is done by choosing the smallest p-value from the

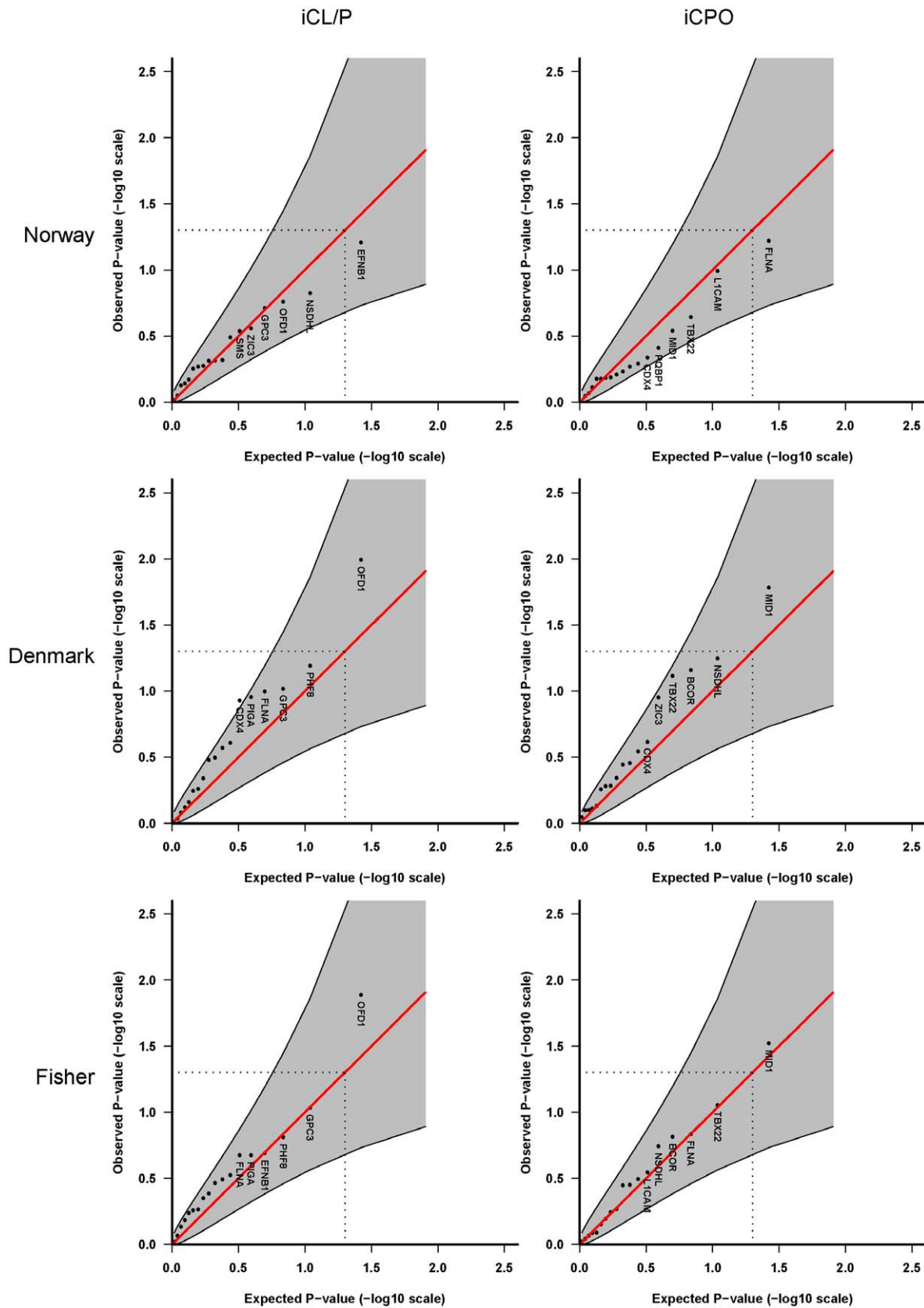
series of windows. If the tests from each window were independent, this p-value could be adjusted with a standard Bonferroni or Šidák correction. However, when analyzing SNPs in strong LD, and in particular when analyzing overlapping windows of length four haplotypes (which share three SNPs with the neighboring haplotype), there is a strong correlation between results obtained from nearby SNPs or windows, and a Bonferroni correction would be far too strict. The `suest` function in HAPLIN corrects the minimum p-value for the dependencies in an optimal manner. This is achieved by first saving the individual (family) score contributions from each window estimation. Under the null hypothesis of no disease association with the haplotypes within a window, the score contributions in that window follow an approximate multivariate normal distribution with mean zero and a covariance matrix which can be computed from the estimated individual score contributions. This allows computing the standard score p-value for that window using a chi-squared test statistic [42]. Then, over a series of windows, the combined score contributions from each window follow approximately a multivariate normal distribution with mean zero and an (empirical) covariance matrix computed from the combined score values. This allows computing the theoretical null distribution of the minimum p-value and thus in calibrating (correcting) the observed minimum p-value. This approach is closely related to the principle of “seemingly unrelated estimation”, as implemented in the statistical package Stata [43].

For the second part, QQ plots were used to inspect visually whether our analyses produced more significant results than what would be expected by chance. The rationale behind the procedure is that if no genes have an effect, the computed p-values should derive from a uniform distribution and thus follow the straight diagonal in the QQ plot. If some genes have an effect (with correspondingly lower p-values), their p-values are likely to show up in the QQ plot as a clear deviation from the diagonal, exhibiting a higher significance than expected by chance under the null hypothesis. We generated QQ plots for each cleft type (iCL/P and iCPO) after combining p-values from the Norwegian and Danish HAPLIN analyses using Fisher’s method [44]. The `pQQ` function in HAPLIN includes confidence bands to assess the size of any deviations from the diagonal. The confidence bands are computed under the null hypothesis of no association between genes and disease. In that case, the order statistic of the sorted p-values will follow a beta distribution with parameters determined by the number of assessed p-values, and the 2.5 and 97.5 percentiles from the corresponding beta distribution provide lower and upper limits, respectively.

**Software.** HAPLIN version 4.1 is implemented in the publicly available **R** statistical package [45] and is freely downloadable from our web site at <http://www.uib.no/smis/gjessing/genetics/software/haplin>. A user-friendly graphical user interface (GUI), which includes some (but not all) of the HAPLIN functionalities, is also freely available at <http://haplin.fhi.no>.

## Study Approval

Clinicopathological information from all participating families and biologic specimens for DNA extraction were obtained with the written informed consent of the mothers and fathers. The study



**Figure 2. Haplotype analyses using up to 4 SNPs per sliding-window, Model 2.**  
doi:10.1371/journal.pone.0039240.g002

was approved by the Norwegian Data Inspectorate, the Regional Committee on Research Ethics for Western Norway, and the respective Institutional Review Boards of the US National Institute of Environmental Health Sciences (NIH/NIEHS) and the University of Iowa. For the Danish orofacial clefts study, study approval was obtained from the regional scientific-ethical committee. All aspects of this research are in compliance with the tenets of the *Declaration of Helsinki* for human research (<http://www.wma.net>).

## Results

**Figures 1 and 2** display the QQ plots for the analysis of 48 SNPs in 18 cleft candidate genes on the X chromosome, without stratification by sex. We first tested a multiplicative model assuming the same relative risk for males and females and no X-inactivation (**Model 2** in **Table 3**). The analyses in **Figure 1** were performed one SNP at a time, whereas **Figure 2** shows the results of the sliding-window haplotype analysis of up to 4 SNPs per window. Overall, the QQ plots show only weak evidence of an association with the oral-facial-digital syndrome 1 gene (*OFD1*, formerly known as *CXORF5*) on chr Xp22 in the Danish iCL/P samples. This association was not replicated in the Norwegian iCL/P samples. For iCPO, there was no evidence of association in either population.

To assess for different gene-effects in males versus females, with effects more evident in one sex stratum than the other, we repeated the analyses for males and females separately. **Figures 3 and 4** show the results of haplotype analysis using a sliding-window of up to 4 SNPs. These analyses correspond to **Model 3** in **Table 3**, in which we assume a multiplicative model with different baseline risks for males and females, different relative risks for males and females, and no X-inactivation. The corresponding single-marker analyses for female and male cases are provided in the online **Figures S1 and S2**, respectively. The association with *OFD1* is now only observed in the Danish iCL/P males, with a relative risk of 2.2 (95% confidence interval: 1.3–3.7; p-value:  $3.6 \times 10^{-3}$ ) with one copy of the variant (minor) allele at the *OFD1* SNP rs2285635 when compared with the reference (major) allele.

Lastly, we tested **Model 4** in which we assume different baseline risks for males and females, a shared relative risk for males and females, and X-inactivation in females. The results of haplotype analysis are depicted in **Figure 5** and the corresponding single-marker analyses are shown in the online **Figure S3**. Again, the only notable association is with *OFD1* in the Danish iCL/P sample only.

## Discussion

Our study was strongly motivated by the unequal sex distribution observed in the two main types of clefts (CL/P and CPO) as well as previous findings of a strong link between X-linked genes and orofacial clefts. X-linked genes have been identified primarily in syndromic forms of clefting and include midline 1 (*MIDI*) on chr Xp22, T-box 22 (*TBX22*) on chr Xq21.1, PHD finger protein 8 (*PHF8*) on chr Xp11.22, and RNA binding motif protein 10 (*RBM10*) on chr Xp11.23. Mutations in *MIDI* cause the X-linked Opitz GBBB syndrome (OSX, MIM 300000), a congenital midline malformation syndrome characterized by clefting of the lip/palate and a variety of other pathologies [46]. An association between specific haplotypes in *MIDI* and isolated CL/P was later reported in an Italian population [19]. Mutations in *TBX22* cause the rare X-linked syndrome ‘cleft palate with ankyloglossia’ (CPX; MIM 303400) [47]. *TBX22* belongs to the T-

box family of genes that are evolutionarily highly conserved and recognized for playing key roles in early vertebrate development. Consistent with the CPX phenotype in humans [47–49], the expression of *Tbx22* in mice is localized to the developing palatal shelves and the base of the tongue. Further, a genome-wide linkage analysis of families with iCL/P identified a susceptibility locus near *TBX22*, suggesting that the linkage signal may emanate from this gene [50]. Mutations in *TBX22* have also been identified in patients with isolated CPO [51,52]. As to *PHF8*, mutations in this gene cause the X-linked mental retardation syndrome Siderius that includes cleft palate as a common phenotypic feature [53,54]. *PHF8* is a histone lysine transcription activator expected to have a wide range of functions. Finally, deep sequencing of exons on the X chromosome identified *RBM10* as the gene causing TARP (MIM 311900), a syndromic form of cleft palate [55].

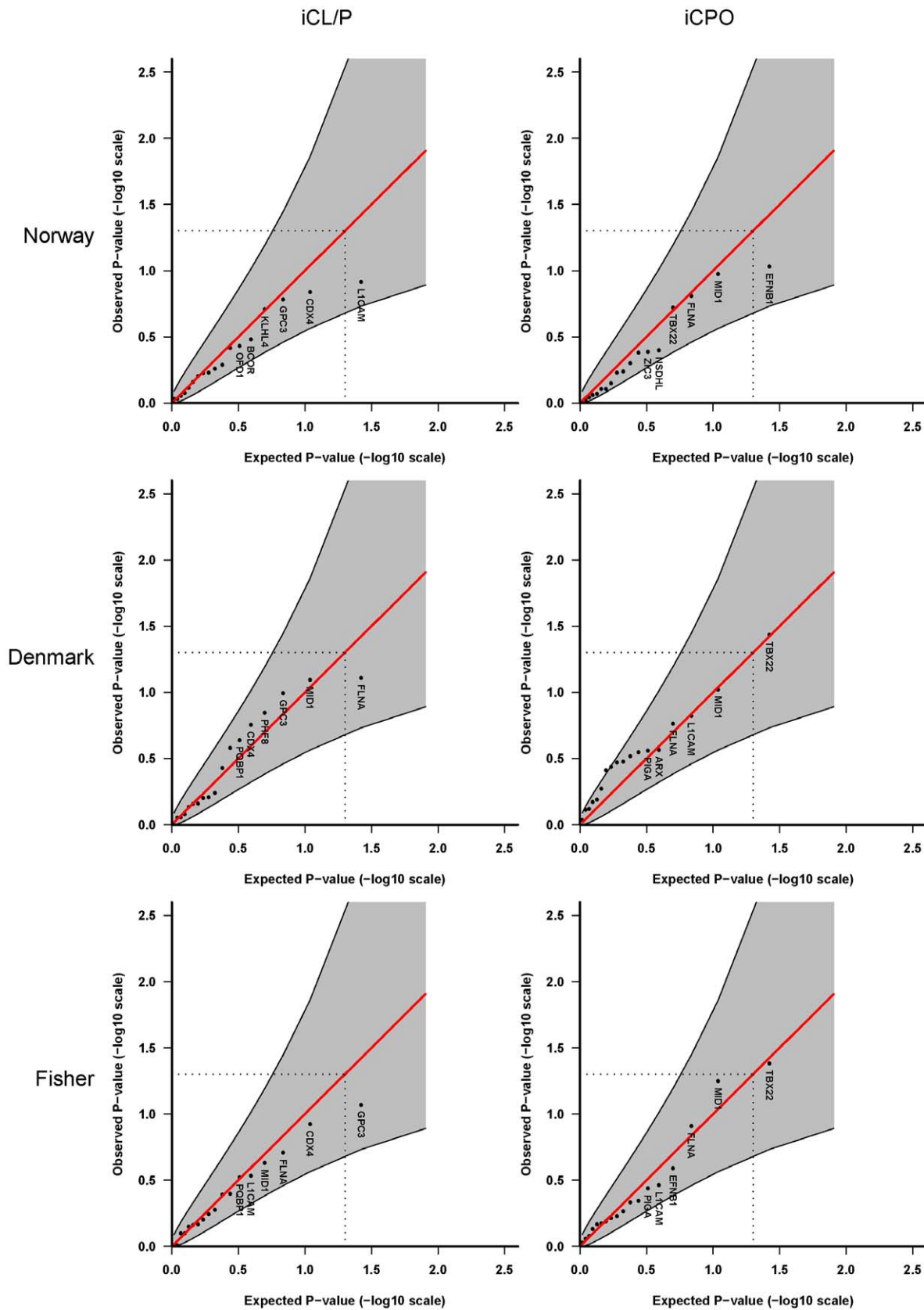
Given these strong links between X-linked genes and syndromic clefts, we examined whether variants in X-linked genes might also be relevant for isolated forms of clefting. To enable X-linked gene analysis, we first developed a method that can i) perform both single-marker and haplotype analyses, ii) generate relevant relative risk estimates with confidence intervals, and iii) assess several etiological models relevant to an X-linked disease locus. The higher prevalence/penetrance for CL/P in males compared with females may be due to hemizyosity for an X-linked disease locus [19]. Therefore, we first analyzed males and females together to account for the possibility that an X-linked disease locus might contribute to clefting risk in both sexes, followed by sex-stratified analyses to investigate whether the X-linked disease locus affects one sex in particular. X-chromosome inactivation in females was also taken into account in the models by treating a heterozygous female ( $X_1X_2$ ) as the average of the two homozygotes ( $X_1X_1$  and  $X_2X_2$ ).

Overall we found only weak associations with *OFD1* in the Danish iCL/P sample, with no replication in the Norwegian iCL/P sample. As noted in our previous analyses of fetal gene-effects in the same study samples [32], the genotype call rates for the Norwegian sample (DNA extracted from blood) and Danish sample (DNA extracted from buccal swabs) were 99.6% and 99.1% respectively. Hence, the lack of replication of the *OFD1* association in the Norwegian iCL/P samples cannot be ascribed to differences in DNA source. Moreover, different genotype frequencies do not imply differences in gene effects on the phenotype.

In sex-stratified analyses, the association of *OFD1* in the Danish iCL/P sample was confined to males only, suggesting a possible sex-specific effect as previously reported for several loci when only males were analyzed [56]. Separate analyses for males and females can be potentially more powerful than a pooled analysis if the X-linked disease locus affects only one sex [25]. An alternative explanation for the apparent sex-specific effect in our data is the potentially higher statistical power to detect an effect of *OFD1* in males due to the larger number of male iCL/P cases available for analysis.

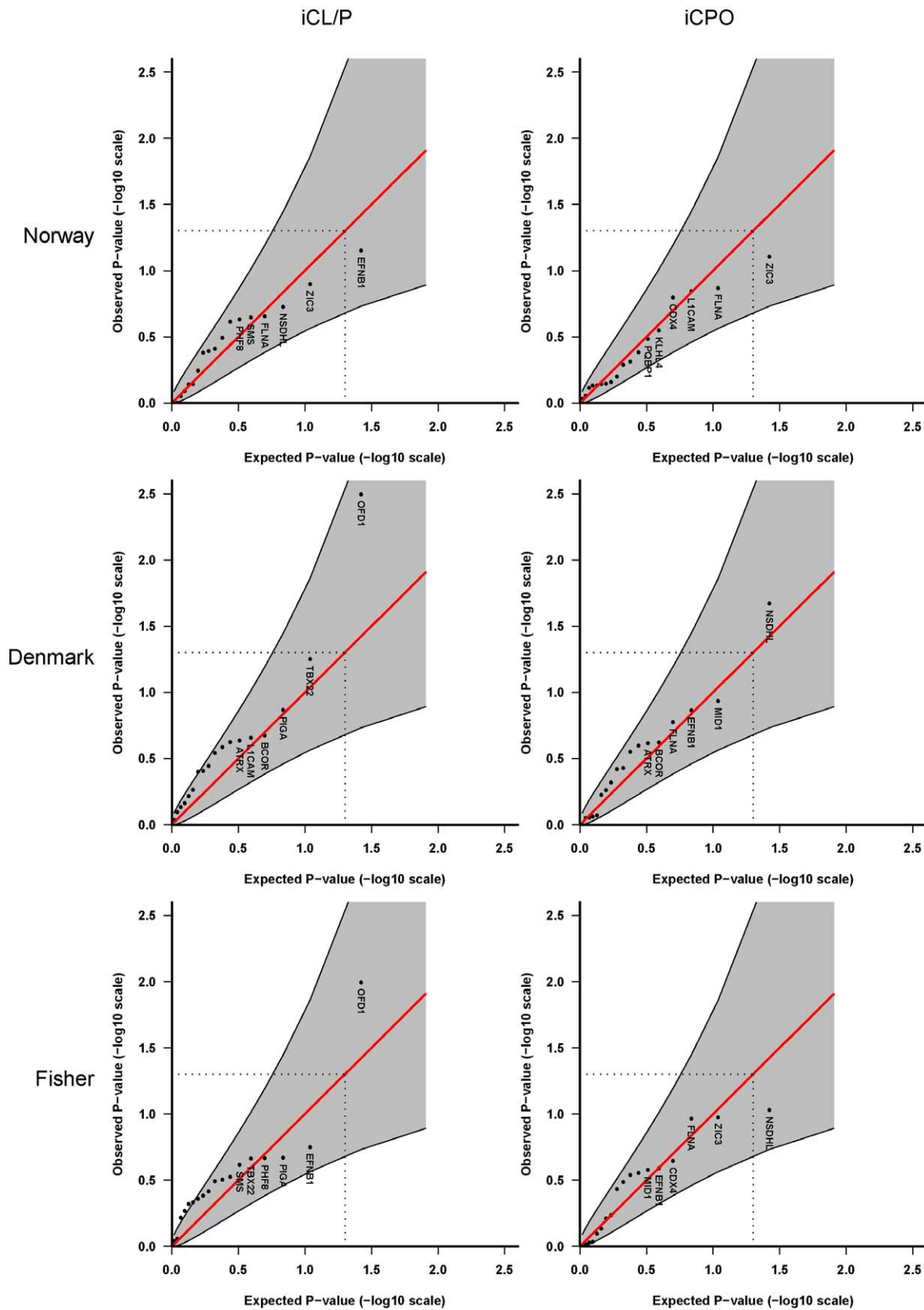
Mutations in *OFD1* underlie the X-linked dominant oral-facial-digital syndrome type 1 (OFD1, MIM 311200), which is characterized by malformations of the face, oral cavity and digits, as well as lethality in the vast majority of affected males [57]. Featuring prominently among the orofacial abnormalities are median cleft lip, clefts of alveolar ridge at the area of lateral incisors, and cleft palate [58]. To our knowledge, however, no genetic association with this gene has previously been reported in isolated clefts.

For haplotype analysis of X-chromosome markers, the standard log-linear approach needs some modification. First, many diseases



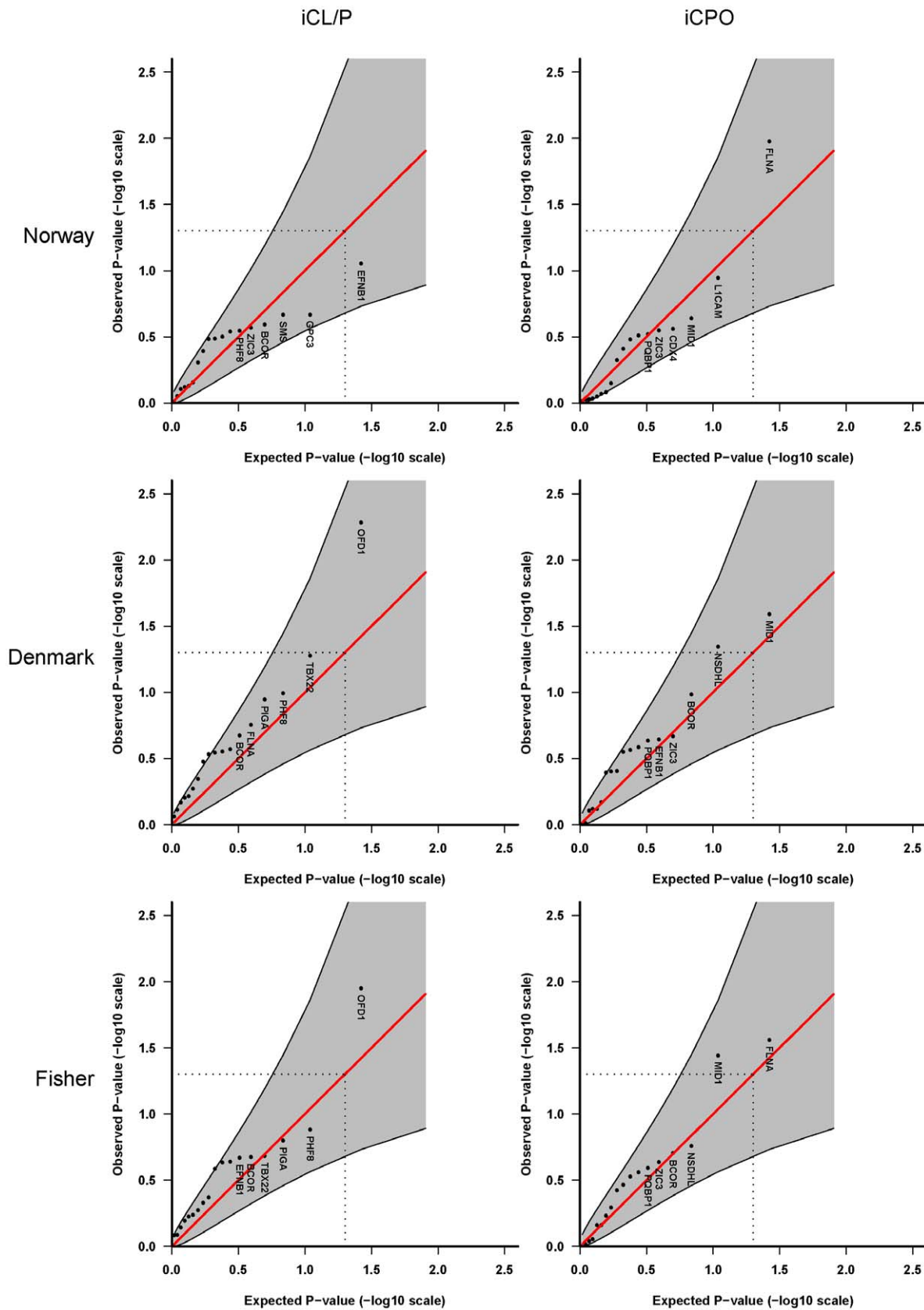
**Figure 3. Haplotype analyses of female cases only using up to 4 SNPs per sliding-window.** These sex-specific analyses are based on **Model 3** in which we assume different baseline risks for males and females, different relative risks for males and females, and no X-inactivation. doi:10.1371/journal.pone.0039240.g003





**Figure 4. Haplotype analyses of male cases only using up to 4 SNPs per sliding-window, Model 3.**

doi:10.1371/journal.pone.0039240.g004



**Figure 5. Haplotype analyses using up to 4 SNPs per sliding-window and taking X-inactivation into account.** These analyses are based on **Model 4** in which we assume different baseline risks for males and females, a shared relative risk for males and females, and X-inactivation. doi:10.1371/journal.pone.0039240.g005

show markedly different birth prevalences in males versus females, as is the case for orofacial clefts, with higher prevalence of CL/P in males and higher prevalence of CPO in females [59]. This difference may be due to causes other than the effect of the particular locus under study, such as loci differentially expressed between males and females. To avoid confounding of the genetic risk estimation by the sex effect, separate baseline risks should be assumed for males and females; i.e. in males the effect of an allele  $A_2$  should be measured relative to the reference allele  $A_1$  in males, whereas in females the effects of  $A_1A_2$  and  $A_2A_2$  should be measured relative to  $A_1A_1$  in females (**Table 3**). Second, it is not clear *a priori* whether a single dose of  $A_2$  in males has an effect comparable to a single dose in females ( $A_1A_2$ ) or to a double dose ( $A_2A_2$ ), or is entirely different from the effect in females. This is aptly illustrated by craniofrontonasal syndrome (CFNS; MIM 304110), an X-linked developmental disorder that paradoxically affects heterozygous females more severely than hemizygous males [60]. In addition, there is the usual question of the relationship between  $A_1A_2$  and  $A_2A_2$  in females; i.e. whether there is a dose-response relationship, a dominant relationship etc. Third, the basic log-linear model in HAPLIN assumes the same allele frequencies for males and females in the background population. While this is a relatively robust assumption for autosomal markers, it is less obvious for X-linked markers. For populations that are genetically relatively homogeneous (like the Danes and Norwegians), however, this assumption seems to be reasonable.

The most extreme solution to the problems raised above is to run separate analyses on males and females. HAPLIN has a special option for doing this, which allows several different response patterns in females to be explored, whereas males are obviously restricted to single-dose effects. To increase statistical power, HAPLIN allows joint analyses of males and females, which reduce the number of parameters to be estimated. The analyses all assume the same allele/haplotype frequencies but different baseline risks for males and females. In addition, various response patterns can be specified.

Another important consideration is X-inactivation in females which may produce a special relationship between male and female allele effects. In females, one X allele in each cell is inactivated (except for a very few second X chromosomes that escape inactivation). A deleterious X-linked allele would be expected to be more detrimental to males than females because males have no chance of compensation by a corresponding normal allele [61]. Because X-inactivation in women occurs in early embryogenesis, women will tend to have a mixture of cells expressing either their mother's or father's X-linked genes (mosaicism). This heterogeneity can have different consequences on a female's disease response depending on how the two X chromosomes are distributed among tissues [61]. The normal expectation would be an equal distribution of the two cell types [61]. However, there may be "founder" effects due to the relatively small number of cells in the embryo at the time of X-inactivation, or differential cellular reproduction rates, leading to an imbalance between the two cell types.

If the risk associated with allele  $A_2$  in females is  $RR_F$ , genotypes  $A_1A_1$  and  $A_2A_2$  will produce risks  $B_F$  and  $B_FRR_F$ , respectively. Assuming a 50:50 cell type distribution, the risk associated with genotype  $A_1A_2$  will be an average of the two homozygotes, i.e.  $(B_F+B_FRR_F)/2 = B_F(1+RR_F)/2$  (**Model 4** in **Table 3**). Techni-

cally speaking this is not a log-linear model, so HAPLIN replaces the heterozygous risk with  $B_F\sqrt{RR_F}$ —the geometric mean of the two homozygous risks. This results in a log-linear model, and as long as  $RR_F$  is neither very small nor very large, the approximation is reasonable. For males, the single-dose effect is then assumed equal to female homozygotes, i.e.  $RR_M = RR_F$  (denoted simply as  $RR$  in **Model 4**). HAPLIN also provides an extension of this model to accommodate an unbalanced cell type distribution.

The basic likelihood models in X-LRT [27] and HAPLIN are similar; for a single SNP, X-LRT uses zero-dose males as reference and estimates relative risks for single-dose males, and zero-, single-, and double-dose females independently. This corresponds to **Model 5** in HAPLIN, and in this special case the results are identical, except that HAPLIN chooses reference levels differently. In addition, HAPLIN provides a number of other modeling options on the X-chromosome, and the software provides a full framework for autosomal and X-linked haplotype association analyses in a candidate-gene or GWAS setting.

To summarize, this is the first candidate-gene based study to investigate the role of X-linked genes in orofacial clefting. Although *OFD1* is a highly plausible gene for clefts, the lack of replication in the Norwegian iCL/P sample highlights the need to confirm these preliminary findings in other datasets. The novel methods presented here address several scenarios relevant to an X-disease locus and can easily be adapted to explore the role of X-linked genes in other complex disorders.

## Supporting Information

### Figure S1 Single-marker analyses of female cases only.

These sex-specific analyses are based on **Model 3** in which we assume different baseline risks for males and females, different relative risks for males and females, and no X-inactivation. (TIF)

### Figure S2 Single-marker analyses of male cases only, Model 3.

(TIF)

### Figure S3 Single-marker analyses taking X-inactivation into account.

These analyses are based on **Model 4** in which we assume different baseline risks for males and females, a shared relative risk for males and females, and X-inactivation. (TIF)

### Table S1 Summary of the 18 X-linked cleft candidate genes and 48 SNPs analyzed in this study.

(DOC)

## Acknowledgments

We are grateful to all participating families for making this study possible. We also thank Drs. Abbe L. Boyles, Min Shi and Jennifer R. Harris for their valuable comments on earlier drafts of this manuscript.

## Author Contributions

Conceived and designed the experiments: AJ ØS RTL AJW KC JCM HKG. Performed the experiments: AJ ØS HKG. Analyzed the data: AJ ØS HKG. Contributed reagents/materials/analysis tools: AJ ØS LC HKG RTL AJW JCM TTN. Wrote the paper: AJ ØS HKG AJW JCM.

## References

1. Strauss RP (1999) The organization and delivery of craniofacial health services: the state of the art. *Cleft Palate Craniofac J* 36: 189–195.
2. Wehby GL, Cassell CH (2010) The impact of orofacial clefts on quality of life and healthcare use and costs. *Oral Dis* 16: 3–10.

3. Dixon MJ, Marazita ML, Beaty TH, Murray JC (2011) Cleft lip and palate: understanding genetic and environmental influences. *Nature reviews Genetics* 12: 167–178.
4. Rahimov F, Jugessur A, Murray JC (2012) Genetics of nonsyndromic orofacial clefts. *Cleft Palate Craniofac J* 49: 73–91.
5. Grosen D, Chevrier C, Skytthe A, Bille C, Molsted K, et al. (2010) A cohort study of recurrence patterns among more than 54,000 relatives of oral cleft cases in Denmark: support for the multifactorial threshold model of inheritance. *Journal of medical genetics* 47: 162–168.
6. Sivertsen A, Wilcox AJ, Skjaerven R, Vindenes HA, Abyholm F, et al. (2008) Familial risk of oral clefts by morphological type and severity: population based cohort study of first degree relatives. *Bmj* 336: 432–434.
7. Grosen D, Bille C, Petersen I, Skytthe A, Hjelmborg JB, et al. (2011) Risk of oral clefts in twins. *Epidemiology* 22: 313–319.
8. Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, et al. (2010) A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nat Genet* 42: 525–529.
9. Grant SF, Wang K, Zhang H, Glaberson W, Annaiah K, et al. (2009) A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. *J Pediatr* 155: 909–913.
10. Mangold E, Ludwig KU, Birnbaum S, Baluado C, Ferrian M, et al. (2010) Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet* 42: 24–26.
11. Kondo S, Schutte BC, Richardson RJ, Bjork BC, Knight AS, et al. (2002) Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat Genet* 32: 285–289.
12. Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, et al. (2008) Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet* 40: 1341–1347.
13. Zuccherro TM, Cooper ME, Maher BS, Daack-Hirsch S, Nepomuceno B, et al. (2004) Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate. *N Engl J Med* 351: 769–780.
14. Liu J, Nyholt DR, Magnussen P, Parano E, Pavone P, et al. (2001) A genome-wide screen for autism susceptibility loci. *American journal of human genetics* 69: 327–340.
15. Lu L, Sun J, Isaacs SD, Wiley KE, Smith S, et al. (2009) Fine-mapping and family-based association analyses of prostate cancer risk variants at Xp11. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 18: 2132–2136.
16. Ohman M, Oksanen L, Kaprio J, Koskenvuo M, Mustajoki P, et al. (2000) Genome-wide scan of obesity in Finnish sibpairs reveals linkage to chromosome Xq24. *The Journal of clinical endocrinology and metabolism* 85: 3183–3190.
17. Pankratz N, Nichols WC, Uniacke SK, Halter C, Murrell J, et al. (2003) Genome-wide linkage analysis and evidence of gene-by-gene interactions in a sample of 362 multiplex Parkinson disease families. *Human molecular genetics* 12: 2599–2608.
18. Burkhardt J, Petit-Teixeira E, Teixeira VH, Kirsten H, Garnier S, et al. (2009) Association of the X-chromosomal genes TIMP1 and IL9R with rheumatoid arthritis. *The Journal of rheumatology* 36: 2149–2157.
19. Scapoli L, Martinelli M, Arlotti M, Palmieri A, Masiero E, et al. (2008) Genes causing clefting syndromes as candidates for non-syndromic cleft lip with or without cleft palate: a family-based association study. *Eur J Oral Sci* 116: 507–511.
20. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52: 506–516.
21. Horvath S, Laird NM, Knapp M (2000) The transmission/disequilibrium test and parental-genotype reconstruction for X-chromosomal markers. *Am J Hum Genet* 66: 1161–1167.
22. Knapp M (2001) Reconstructing parental genotypes when testing for linkage in the presence of association. *Theoretical population biology* 60: 141–148.
23. Ho GY, Bailey-Wilson JE (2000) The transmission/disequilibrium test for linkage on the X chromosome. *Am J Hum Genet* 66: 1158–1160.
24. Ding J, Lin S, Liu Y (2006) Monte Carlo pedigree disequilibrium test for markers on the X chromosome. *Am J Hum Genet* 79: 567–573.
25. Chung RH, Morris RW, Zhang L, Li YJ, Martin ER (2007) X-APL: an improved family-based test of association in the presence of linkage for the X chromosome. *Am J Hum Genet* 80: 59–68.
26. Zhang L, Martin ER, Morris RW, Li YJ (2009) Association test for X-linked QTL in family-based designs. *Am J Hum Genet* 84: 431–444.
27. Zhang L, Martin ER, Chung RH, Li YJ, Morris RW (2008) X-LRT: a likelihood approach to estimate genetic risks and test association with X-linked markers using a case-parents design. *Genet Epidemiol* 32: 370–380.
28. Gjessing HK, Lie RT (2006) Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. *Ann Hum Genet* 70: 382–396.
29. Migeon BR (2011) The single active X in human cells: evolutionary tinkering personified. *Hum Genet* 130: 281–293.
30. Bille C, Knudsen LB, Christensen K (2005) Changing lifestyles and oral clefts occurrence in Denmark. *Cleft Palate Craniofac J* 42: 255–259.
31. Wilcox AJ, Lie RT, Solvoll K, Taylor J, McConaughy DR, et al. (2007) Folic acid supplements and risk of facial clefts: national population based case-control study. *Bmj* 334: 464.
32. Jugessur A, Shi M, Gjessing HK, Lie RT, Wilcox AJ, et al. (2009) Genetic determinants of facial clefting: analysis of 357 candidate genes using two national cleft studies from Scandinavia. *PLoS ONE* 4: e5385.
33. Umbach DM, Weinberg CR (1997) Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat Med* 16: 1731–1743.
34. Umbach DM, Weinberg CR (2000) The use of case-parent triads to study joint effects of genotype and exposure. *Am J Hum Genet* 66: 251–261.
35. Weinberg CR (1999) Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am J Hum Genet* 65: 229–235.
36. Weinberg CR (1999) Allowing for missing parents in genetic studies of case-parent triads. *Am J Hum Genet* 64: 1186–1193.
37. Weinberg CR, Morris RW (2003) Invited commentary: Testing for Hardy-Weinberg disequilibrium using a genome single-nucleotide polymorphism scan based on cases only. *Am J Epidemiol* 158: 401–403; discussion 404–405.
38. Weinberg CR, Umbach DM (2000) Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *Am J Epidemiol* 152: 197–203.
39. Weinberg CR, Umbach DM (2005) A hybrid design for studying genetic influences on risk of diseases with onset early in life. *Am J Hum Genet* 77: 627–636.
40. Weinberg CR, Wilcox AJ, Lie RT (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet* 62: 969–978.
41. Wilcox AJ, Weinberg CR, Lie RT (1998) Distinguishing the effects of maternal and offspring genes through studies of “case-parent triads”. *Am J Epidemiol* 148: 893–901.
42. Pawitan Y (2001) In all likelihood. Oxford: Clarendon Press.
43. StataCorp (2009) Stata: Release 11. Statistical Software. College Station, Texas, USA: Stata Press.
44. Fisher R (1958) Statistical Methods for Research Workers. New York: Hafner.
45. R Development Core Team (2006) R: A Language and Environment for Statistical Computing.
46. Quaderi NA, Schweiger S, Gaudenz K, Franco B, Rugarli EI, et al. (1997) Opitz G/BBB syndrome, a defect of midline development, is due to mutations in a new RING finger gene on Xp22. *Nat Genet* 17: 285–291.
47. Braybrook C, Doudney K, Marciano AC, Arnason A, Bjornsson A, et al. (2001) The T-box transcription factor gene TBX22 is mutated in X-linked cleft palate and ankyloglossia. *Nat Genet* 29: 179–183.
48. Braybrook C, Ligo S, Doudney K, Henderson D, Marciano AC, et al. (2002) Craniofacial expression of human and murine TBX22 correlates with the cleft palate and ankyloglossia phenotype observed in CPX patients. *Hum Mol Genet* 11: 2793–2804.
49. Pauws E, Hoshino A, Bentley L, Prajapati S, Keller C, et al. (2009) Tbx22null mice have a submucous cleft palate due to reduced palatal bone formation and also display ankyloglossia and choanal atresia phenotypes. *Hum Mol Genet* 18: 4171–4179.
50. Prescott NJ, Lees MM, Winter RM, Malcolm S (2000) Identification of susceptibility loci for nonsyndromic cleft lip with or without cleft palate in a two stage genome scan of affected sib-pairs. *Hum Genet* 106: 345–350.
51. Marciano AC, Doudney K, Braybrook C, Squires R, Patton MA, et al. (2004) TBX22 mutations are a frequent cause of cleft palate. *J Med Genet* 41: 68–74.
52. Suphaeetiporn K, Tongkobpetch S, Siriwan P, Shotelersuk V (2007) TBX22 mutations are a frequent cause of non-syndromic cleft palate in the Thai population. *Clin Genet* 72: 478–483.
53. Abidi FE, Miano MG, Murray JC, Schwartz CE (2007) A novel mutation in the PHF8 gene is associated with X-linked mental retardation with cleft lip/cleft palate. *Clin Genet* 72: 19–22.
54. Laumonnier F, Holbert S, Ronce N, Faravelli F, Lenzner S, et al. (2005) Mutations in PHF8 are associated with X linked mental retardation and cleft lip/cleft palate. *J Med Genet* 42: 780–786.
55. Johnston JJ, Teer JK, Cherukuri PF, Hansen NF, Loftus SK, et al. (2010) Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am J Hum Genet* 86: 743–748.
56. Mangold E, Reutter H, Birnbaum S, Walier M, Mattheisen M, et al. (2009) Genome-wide linkage scan of nonsyndromic orofacial clefting in 91 families of central European origin. *Am J Med Genet A* 149A: 2680–2694.
57. Ferrante MI, Giorgio G, Feather SA, Bulfone A, Wright V, et al. (2001) Identification of the gene for oral-facial-digital type I syndrome. *Am J Hum Genet* 68: 569–576.
58. Jones KL (2006) Oral-Facial-Digital Syndrome (OFD Syndrome, Type 1). *SMITH'S Recognizable Patterns of Human Malformation*. 6th Edition ed: Elsevier Inc.
59. Mossey PA, Little J (2002) Epidemiology of oral clefts: an international perspective. In: Wyszynski DFE, editor. *Cleft lip and palate: from origin to treatment*. New York: Oxford University Press. 127–158.
60. Twigg SR, Kan R, Babbs C, Bochkova EG, Robertson SP, et al. (2004) Mutations of ephrin-B1 (EFNB1), a marker of tissue boundary formation, cause craniofrontonasal syndrome. *Proc Natl Acad Sci U S A* 101: 8652–8657.
61. Migeon BR (2006) The role of X inactivation and cellular mosaicism in women's health and sex-specific diseases. *JAMA* 295: 1428–1433.